BJTU's Submission to CCMT 2021 Translation Evaluation Task

Xiang Li, Xuanxuan Wu, Shuaibo Wang, Xiaoke Liang, Li Yin, Hui Huang,
Yufeng Chen, and Jin'an Xu(✉)

School of Computer and Information Technology, Beijing Jiaotong University,
Beijing, China
{xiang.li,19120414,20120419,20125185,20125265,18112023,chenyf,
jaxu}@bjtu.edu.cn

Abstract. This paper presents the systems developed by Beijing Jiaotong University for the CCMT2021 evaluation tasks. We joined four translation tasks of Chinese-English, English-Chinese, Uyghur-Chinese, Tibetan-Chinese. In all directions, we build our system based on transformer architecture and Dynamic-Conv. Additionally, we apply Byte Pair Encoding (BPE) to all translation tasks to resolve the out-of-vocabulary (OOV) problem. We also adopt some techniques that have been proven effective recently in academia, such as data augmentation, finetuning, model ensemble and reranking. Experiments show that our machine translation systems achieved high accuracy on all directions.

Keywords: Neural machine translation · Data augmentation · Finetuning · Model ensemble · Reranking

1 Introduction

This paper introduces in detail the submission of Beijing Jiaotong University to the translation evaluation task in the 17-th China Conference on Machine Translation (CCMT2021). We participated in both directions of Chinese-English translation tasks from the news field and two minority language translation tasks Tibetan-Chinese translation from government literature and Uyghur-Chinese translation from the news field.

In these directions, we built our system based on five different architectures, the first one is solely based on attention mechanisms, namely the Transformer-base model [11]. We broadened Transformer with bigger hidden dimensions and more attention headers to better extract features from source segments, which is named as Transformer-big. We also tried to augmented the encoder layers to extract more semantic information from the source which is named as Transformer-deep. Transformer-big and Transformer-deep are proved to outperform Transformer-base model in most cases [12]. Additionally, we also tried to substitute the self-attention layer with lightweight convolution, providing us with another different model to use when doing model ensemble [14].

J. Su and R. Sennrich (Eds.): CCMT 2021, CCIS 1464, pp. 117–124, 2021.
https://doi.org/10.1007/978-981-16-7512-6_10

Additionally, we applied sub-word segmentation to both languages to resolve the out-of-vocabulary problem [9]. To deal with the scarcity of training data, we created massive synthetic data using back translation based on monolingual Chinese data [8]. To make full use of the translation knowledge learned by other decoding models, knowledge distillation is used to integrate various knowledge into one model [3].

The in-domain finetuning is very effective in our experiments and especially, we used a boosted finetuning method for Chinese→English and English→ Chinese tasks. We also take advantage of the combination methods to further improve the translation quality.

We also applied two model ensemble techniques, namely model averaging and model ensemble, to leverage multiple models to further improve the result [1,7]. To alleviate unbalanced output and error accumulation during left to right decoding, we performed reranking on the top-k outputs based on z-Mert algorithm [4].

2 Data

2.1 Chinese-English

We use all available data provided by CCMT'21 and WMT'21, which contain 28.6M bilingual sentence pairs and 100M Chinese Monolingual data and 120M English Monolingual data. We apply the following procedures to preprocess the data:

1. Remove illegal UTF-8 characters and replace control characters with a single space.
2. Convert Traditional Chinese sentences into Simplified Chinese.
3. Apply Unicode NFKC normalization.
4. Remove duplicated sentence pairs.
5. Keep parallel sentences with a length ratio between 0.7-2.2.
6. Truecase[1] the English corpus.

For the new corpus "ParaCrawl v7.1" in WMT'21, there are plenty of noisy sentence pairs. We have trained a baseline model with Transformer-base to filter out the noisy pairs with SacreBLEU lower than 35.0.

2.2 Uyghur→Chinese

We use the parallel data provided by CCMT'21, which contains 0.17M pairs. We cleaned the provided training data accords to two criteria, namely the length ratio of source to target for each sentence pair, and the average length of source sentence and target sentence.

[1] https://github.com/moses-smt/mosesdecoder.

2.3 Tibetan→Chinese

We use all available data provided by CCMT'21, which has 0.15M parallel sentences. However we have not used the devset provided by CCMT'21, we randomly sample 1k sentences in parallel sentences as our devset and the rest of the available data is used as training data. We apply Unicode NFKC to normalize the data. For Tibetan word segmentation, we build a vocabulary which consists of 140k words, and use Bidirectional-Maximum Matching algorithm.

3 Model

As we explained before, we combined four different architectures in our work, namely Transformer-base, Transformer-deep, Transformer-big and Light-Conv.

Transformer-Base. Transformer is a completely attention-based structure for dealing with problems related to sequence models [10], such as machine translation. The Transformer model does not use any CNN or RNN structure, capable of working in the process of highly parallelization, so the training speed is very fast while improving the translation performance. Transformer-base is the naive version of transformer.

Transformer-Deep. The performance of Transformer can be improved by increasing the number of layers in the encoder. We follow to use deep Transformer. To address the vanishing-gradient problem in deep Transformer, we use the post-layer normalization instead of the pre-layer normalization. In Chinese-English directions, we adopt this model which has great performance.

Transformer-Big. In some cases, Transformer-deep does not perform better than big, which have a fewer parameters than the former. Therefore, for the stage of training and inference, Transformer is faster than Transformer-deep.

LightConv. Lightweight convolution uses the prototype of deep (separable) convolution in CV domain, which greatly reduces the number of parameters and reduces the complexity by sharing parameters in the channel dimension. On the basis of light weight, dynamic convolution is proposed, where the weight of CNN is calculated dynamically from the input feature. The Dynamic-Conv model is proved to be competitive with Transformer model in many scenarios.

4 Method

4.1 Data Augmentation

Back-Translation. We augment the training data by exploring the monolingual corpus using back translation. Specifically, we select target monolingual corpus which has the same size as the training corpus and then translate them back into the source language using target-to-source (T2S) models. We merge the synthetic data with the bilingual data to train our models. We also add noise to the translated sentences to further improve the performance namely Noisy Back-Translation.

Knowledge Distillation. The existing translation model decodes from left to right (L2R), and from source to target (S2T). In order to make full use of the translation knowledge learned by other decoding models, knowledge distillation is adopted to improve the translation performance. Knowledge distillation is a method for knowledge transfer, where the prediction distribution of teacher model is used to guide the parameter learning of student model. In our submission, the following three teacher models are trained first:

1. The translation model decodes from source to target and from left to right (L2R).
2. The translation model decodes from source to target and from right to left (R2L).
3. The translation model decodes from target to source and from left to right (T2S).

After obtaining the above three translation models, we use the method of sentence level knowledge distillation to decode the training data and get their respective decoding results, and form the bilingual sentence pairs of knowledge distillation with their respective input sentences. In this evaluation, we mixed the knowledge distillation bilingual sentence pairs with the original training data. In this way, in mixed bilingual data, in addition to the original training data, it also contains the prediction results of the respective teacher models. Finally, the student model is retrained with mixed training data.

4.2 Model Average

Because of the mismatch of BLEU and MLE Loss in the final convergence stage, we have applied the Model Average method to average the parameters from the last several checkpoints. We have found that Model Average works on Uyghur→Chinese and Tibetan→Chinese but makes no sense in Chinese-English.

4.3 Finetune

Finetuning [2] with in-domain data can bring huge improvements. We also use development set as the in-domain dataset. The source side of newsdev2017, newstest2017 and newstest2018 are composed of two parts: documents created originally in Chinese and documents created originally in English. We split these datasets into original Chinese part and original English part according to tag attributes of SGM files. For Chinese-English translation, we use CWMT2008, CWMT2009 and original Chinese part of newsdev2017, newstest2017, newstest2018 and newstest2020 as the in-domain dataset. For English-Chinese translation, we use original English part of newsdev2017, newstest2017, newstest2018 and newstest2020 as the in-domain dataset. During finetuning, we use a larger dropout rate, a smaller constant learning rate and batch size. The parameters are updated after each epoch, which is enabled by using gradient accumulation.

4.4 Model Ensemble

Ensemble is a well-known technique to combine different models for stronger performance. We utilize the frequently used method for ensemble, which calculates the word level averaged log-probability among different models during decoding. On account of the model diversity among single models has a strong impact on the performance of ensembling models, we combine single models that have different model architectures (Transformer-base, Transformer-big, Transformer-deep, Transfomer-deepbig, Light-conv, Dynamic-conv). We also try to use Transductive Ensemble Learning (TEL) [13] to replace ensemble. TEL is a technique utilizing the synthetic test data (consists of original source sentences and translations of target-side) of different models to finetune a single model.

4.5 Reranking

Neural machine translation models are usually decoded from left to right, and are faced with the problem of unbalanced output and error accumulation. In the process of translation generation, if there are errors in the first few moments, it is difficult to produce correct results in the following. To some extent, this problem can be alleviated by increasing the space of beam search. However, since we only select the sentence with the highest prediction probability as the final output, the increase of searching space will not bring significant benefits, and even bring some performance losses. Therefore, this paper uses the method of reranking. In this paper, several feature models are trained to grade the candidate translation. The feature models include the R2L model, L2R model, T2S model and language model scores. Word-penalty is also included to penalize too short output, which is the length of each candidate. After that, z-Mert [15] is used to rerank the candidate translations, and the translation with the highest score is selected as the final output translation.

5 Experiment

5.1 Chinese→English

We use the PyTorch implementation of open-source toolkit fairseq [5] to conduct all experiments. To enable open vocabulary, we learn 32K BPE operations separately on Chinese and English texts using subword-nmt toolkit. We set Chinese vocabulary size of 40k and English vocabulary size of 32k. All models are trained on Tesla-V100. Table 1 shows the results of Chinese-English Translation on newstest2019 dataset. All methods we used can bring substantial improvement over the baseline system. Applying data augmentation methods improve the baseline system by 2.3 BLEU score. Finetuning is the most effective approach. With transductive ensemble on newstest2019 our model has achieved 40.12.

Table 1. BLEU evaluation results on the newstest2019 Chinese-English test set

Settings	Transformer-big	Transformer-deep	Lightconv
Baseline	27.72	28.14	27.11
+ data augment	30.12	30.07	29.88
+ finetuning	39.32	38.55	38.32
Ensemble	40.12		

5.2 English→Chinese

We have the same preprocessing setting with the Chinese→English direction. And all the models are trained on RTX 1080Ti. However, the back-translation does not work, therefore we just apply noisy back-translation. Our results are depicted as Table 2 where finetuning in English→Chinese does not have the same improvement.

Table 2. BLEU evaluation results on newstest2019 English-Chinese test set

Settings	Transformer-big	Transformer-deep	Lightconv
Baseline	36.96	35.75	–
+ data augment	37.34	36.66	37.43
+ finetuning	38.47	37.46	38.77
Ensemble	39.81		

5.3 Uyghur→Chinese

In Uyghur, we adopt fast-align and kenLM to select the monolingual data. We then back translate the monolingual sentences to generate the twice size of the parallel data. Finally we combine the parallel data and the pseudo-parallel data.

In our experiment, we utilize the BPE-Dropout [6] as a method of data augmentation with the dropout rate 0.1. BPE-Dropout performs well on the mini-scale dataset.

Table 3. BLEU evaluation results on CCMT'21 Uyghur-Chinese dev set

Settings	Transformer-base	Transformer-big	Transformer-deepbig	Dynamic-conv
Baseline	41.12	40.09	41.03	43.91
+ data augument	43.96	43.85	43.75	45.17
+ finetuning	44.65	45.03	45.79	45.22
Ensemble	48.02			

Table 4. Result of BPE-dropout in Uyghur-Chinese.

Models	BPE	BPE-dropout
Transformer-base	41.12	43.96
Dynamic-conv	43.91	45.17

Table 5. BLEU evaluation results on CCMT'21 Tibetan-Chinese dev set.

Settings	Transformer-base	Transformer-big	Dynamic-conv
Baseline	46.83	46.48	46.76
+ data augument	47.34	46.91	46.97
+ finetuning	48.50	47.62	48.90
Ensemble	51.09		
Rerank	54.35		

5.4 Tibetan→Chinese

Table 5 shows the result of Tibetan→Chinese that reranking has improved 3.3 BLEU score, which does not make sense in Chiense→Englsih, English→Chinese and Uyghur→Chinese.

6 Conclusion

In this paper, we described our submission in four translation evaluation projects including Chinese to English, English to Chinese, Tibetan to Chinese and Uyghur to Chinese. In all directions, we build our system based on six different architectures, namely Transformer-base, Transformer-big, Transformer-deep, Transformer-deepbig and Dynamic-Conv. Finally, we obtain substantial improvements combining these methods. Our training strategies including back-translation, knowledge distillation, model ensemble and reranking have good performance in these tasks.

Acknowledgement. The research work descried in this paper has been supported by the National Key R&D Program of China 2020AAA0108001 and the National Nature Science Foundation of China (No. 61976015, 61976016, 61876198 and 61370130). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

1. Chen, H., Lundberg, S., Lee, S.I.: Checkpoint ensembles: Ensemble methods from a single training process. arXiv preprint arXiv:1710.03282 (2017)
2. Chu, C., Dabre, R., Kurohashi, S.: An empirical comparison of domain adaptation methods for neural machine translation. In: ACL, pp. 385–391 (2017)
3. Kim, Y., Rush, A.M.: Sequence-level knowledge distillation. In: EMNLP, pp. 1317–1327 (2016)
4. Olteanu, M., Suriyentrakorn, P., Moldovan, D.: Language models and reranking for machine translation. In: Proceedings on the Workshop on Statistical Machine Translation, pp. 150–153 (2006)
5. Ott, M., et al.: fairseq: a fast, extensible toolkit for sequence modeling. In: NAACL, pp. 48–53 (2019)
6. Provilkov, I., Emelianenko, D., Voita, E.: Bpe-dropout: Simple and effective subword regularization. arXiv preprint arXiv:1910.13267 (2019)
7. Rokach, L.: Ensemble-based classifiers. Artif. Intell. Rev. **33**(1–2), 1–39 (2010). https://doi.org/10.1007/s10462-009-9124-7
8. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: ACL, pp. 86–96 (2016)
9. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: ACL, pp. 1715–1725 (2016)
10. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS, pp. 3104–3112 (2014)
11. Vaswani, A., Shazeer, N., et al.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)
12. Wang, Q., et al.: Learning deep transformer models for machine translation. arXiv preprint arXiv:1906.01787 (2019)
13. Wang, Y., Wu, L., Xia, Y., Qin, T., Zhai, C., Liu, T.Y.: Transductive ensemble learning for neural machine translation. In: AAAI, pp. 6291–6298 (2020)
14. Wu, F., Fan, A., Baevski, A., Dauphin, Y.N., Auli, M.: Pay less attention with lightweight and dynamic convolutions. arXiv preprint arXiv:1901.10430 (2019)
15. Zaidan, O.: Z-MERT: a fully configurable open source tool for minimum error rate training of machine translation systems. Prague Bull. Math. Linguist. **91**, 79–88 (2009)